Estimating the conditional density by histogram type estimators and model selection

Mathieu Sart

ESTIMATING THE CONDITIONAL DENSITY BY HISTOGRAM TYPE ESTIMATORS AND MODEL SELECTION

MATHIEU SART

Abstract. We propose a new estimation procedure of the conditional density for independent and identically distributed data. Our procedure aims at using the data to select a function among arbitrary (at most countable) collections of candidates. By using a deterministic Hellinger distance as loss, we prove that the selected function satisfies a non-asymptotic oracle type inequality under minimal assumptions on the statistical setting. We derive an adaptive piecewise constant estimator on a random partition that achieves the expected rate of convergence over (possibly inhomogeneous and anisotropic) Besov spaces of small regularity. Moreover, we show that this oracle inequality may lead to a general model selection theorem under very mild assumptions on the statistical setting. This theorem guarantees the existence of estimators possessing nice statistical properties under various assumptions on the conditional density (such as smoothness or structural ones).

1. Introduction

Let $(X_i, Y_i)_{1\leq i\leq n}$ be n independent and identically distributed random variables defined on an abstract probability space $(\Omega, \mathcal{E}, \mathbb{P})$ with values in $\mathbb{X}\times\mathbb{Y}$. We suppose that the conditional law $\mathcal{L}(Y_i \mid X_i)$ admits a density $s(X_i, \cdot)$ with respect to a known σ-finite measure μ. In this paper, we address the problem of estimating the conditional density s on a given subset $A \subset \mathbb{X}\times\mathbb{Y}$.

When (X_i, Y_i) admits a joint density $f_{(X,Y)}$ with respect to a product measure $\nu\otimes\mu$, one can rewrite s as

$$s(x,y) = \frac{f_{(X,Y)}(x,y)}{f_X(x)} \quad \text{for all } x, y \in \mathbb{X}\times\mathbb{Y} \text{ such that } f_X(x) > 0,$$

where f_X stands for the density of X_i with respect to ν. A first approach to estimate s was introduced in the late 60's by Rosenblatt (1969). The idea was to replace the numerator and the denominator of this ratio by Kernel estimators. We refer to Hyndman et al. (1996) for a study of its asymptotic properties. An alternative point of view is to consider the conditional estimation density problem as a non-parametric regression problem. This has motivated the definition of local parametric estimators which have been asymptotically studied by Fan et al. (1996); Hyndman and Yao (2002); De Gooijer and Zerom (2003). Another approach was proposed by Faugeras (2009). He showed asymptotic results for his copula based estimator under smoothness assumptions on the marginal density of Y and the copula function.

Date: October, 2016.

2010 *Mathematics Subject Classification.* 62G05, 62G07.

Key words and phrases. Adaptive estimation, Conditional density, Histogram, Model selection, Robust tests.

The aforementioned procedures depend on some parameters that should be tuned according to the (usually unknown) regularity of s. The practical choice of these parameters is for instance discussed in Fan and Yim (2004) (see also the references therein). Nonetheless, adaptive estimation procedures are rather scarce in the literature. We can cite the procedure of Efromovich (2007) which yields an oracle inequality for an integrated $\mathbb{L}^2$ loss. His estimator is sharp minimax under Sobolev type constraints. Bott and Kohler (2015) adapted the combinatorial method of Devroye and Lugosi (1996) to the problem of bandwidth selection in kernel conditional density estimation. They showed that this method allows to select the bandwidth according to the regularity of s by proving an oracle inequality for an integrated $\mathbb{L}^1$ loss. The papers of Brunel et al. (2007); Akakpo and Lacour (2011) are based on the minimisation of a penalized $\mathbb{L}^2$ contrast inspired from the least squares. They established a model selection result for an empirical $\mathbb{L}^2$ loss and then for an integrated $\mathbb{L}^2$ loss. These procedures build adaptive estimators that may achieve the minimax rates over Besov classes. The paper of Chagny (2013) is based on projection estimators, Goldenshluger and Lepski methodology and a transformation of the data. She showed an oracle inequality for an integrated $\mathbb{L}^2$ loss from which she deduced that her estimator is adaptive and reaches the expected rate of convergence under Sobolev constraints on an auxiliary function. Cohen and Le Pennec (2011) gave model selection results for the penalized maximum likelihood estimator for a loss based on a Jensen-Kullback-Leibler divergence and under bracketing entropy type assumptions on the models.

Another estimation procedure that can be found in the literature is the one of T-estimation (T for test) developed by Birgé (2006). It leads to much more general model selection theorems, which allows the statistician to model finely the knowledge he has on the target function to obtain accurate estimates. It is shown in Birgé (2012) that one can build a T-estimator of the conditional density. We now define the loss used in that paper to compare it with ours. We suppose henceforth that the distribution of X_i is absolutely continuous with respect to a known σ-finite measure ν. Let f_X be its Radon-Nikodym derivative. We denote by $\mathbb{L}^1_+(A, \nu\otimes\mu)$ the cone of non-negative integrable functions on $\mathbb{X}\times\mathbb{Y}$ with respect to the product measure $\nu\otimes\mu$ vanishing outside A. Birgé (2012) measured the quality of his estimator by means of the Hellinger deterministic distance δ defined by

$$\delta^2(f,g) = \frac{1}{2}\int_A \left(\sqrt{f(x,y)} - \sqrt{g(x,y)}\right)^2 \mathrm{d}\nu(x)\,\mathrm{d}\mu(y) \quad \text{for all } f,g\in\mathbb{L}^1_+(A,\nu\otimes\mu).$$

It is assumed in that paper that the marginal density f_X of X is bounded from below by a positive constant. This classical assumption seems natural in the sense that the estimation of s is better in regions of high value of f_X than regions of low value as stressed for instance in Bertin et al. (2013). In the present paper, we bypass this assumption by measuring the quality of our estimators through the Hellinger distance h defined by

$$h^2(f,g) = \frac{1}{2}\int_A \left(\sqrt{f(x,y)} - \sqrt{g(x,y)}\right)^2 f_X(x)\,\mathrm{d}\nu(x)\,\mathrm{d}\mu(y) \quad \text{for all } f,g\in\mathbb{L}^1_+(A,\nu\otimes\mu).$$

The marginal density f_X can even vanish, in contrast to most of the papers cited above. We propose a new and data-driven (penalized) criterion adapted to this unknown loss. Its definition is in the line of the ideas developed in Baraud (2011); Sart (2014); Baraud et al. (2016).

The main result is an oracle type inequality for (at most) countable families of functions of $\mathbb{L}^1_+(A,\nu\otimes\mu)$. This inequality holds true without additional assumptions on the statistical setting. We use it a first time as an alternative to resampling methods to select among families

of piecewise constant estimators. We deduce an adaptive estimator that achieves the expected rates of convergence over a range of (possibly inhomogeneous and anisotropic) Besov classes, including the ones of small regularities. A second application of this inequality leads to a new general model selection theorem under very mild assumptions on the statistical setting. We propose 3 illustrations of this result. The first shows the existence of an adaptive estimator that attains the expected rate of convergence (up to a logarithmic term) over a very wide range of (possibly inhomogeneous and anisotropic) Besov spaces. This estimator is therefore able to cope in a satisfactory way with very smooth conditional densities as well as with very irregular ones. The second illustration deals with the celebrated regression model. It shows that the rates of convergence can be faster than the ones we would obtain under pure smoothness assumptions on s when the data actually obey to a regression model (not necessarily Gaussian). The last illustration concerns the case where the random variables X_i lie in a high dimensional linear space, say $\mathbb{X} = \mathbb{R}^{d_1}$ with d_1 large. In this case, we explain how our procedure can circumvent the curse of dimensionality.

The paper is organized as follows. In Section 2, we carry out the estimation procedure and the oracle inequality. We use it to select among a family of piecewise constant estimators and study the quality of the selected estimator. Section 3 is dedicated to the general model selection theorem and its applications. The proofs are postponed to Section 4.

We now introduce the notations that will be used all along the paper. We set $\mathbb{N}^\star = \mathbb{N} \setminus \{0\}$, $\mathbb{R}_+ = [0, +\infty)$, $\mathbb{R}_+^\star = (0, +\infty)$. For $x, y \in \mathbb{R}$, $x \wedge y$ (respectively $x \vee y$) stands for $\min(x, y)$ (respectively $\max(x, y)$). The positive part of a real number x is denoted by $x_+ = x \vee 0$. The distance between a point x and a set A in a metric space (E, d) is denoted by $d(x, A) = \inf_{y \in A} d(x, y)$. The cardinality of a finite set A is denoted by $|A|$. The restriction of a function f to a set A is denoted by $f_{|A}$. The indicator function of a set A is denoted by 1_A. The notations $c, C, c', C', c_1, C_1, c_2, C_2, \ldots$ are for the constants. These constants may change from line to line.

2. Selection among points and hold-out

Throughout the paper, $n > 3$ and A is of the form $A = A_1 \times A_2$ with $A_1 \subset \mathbb{X}$, $A_2 \subset \mathbb{Y}$.

2.1. Selection rule and main theorem. Let $\mathcal{L}(A, \mu)$ be the subset of $\mathbb{L}_+^1(A, \nu \otimes \mu)$ defined by

$$\mathcal{L}(A, \mu) = \left\{ f \in \mathbb{L}_+^1(A, \nu \otimes \mu),\ \sup_{x \in A_1} \int_{A_2} f(x, y)\, \mathrm{d}\mu(y) \leq 1 \right\},$$

and let S be an at most countable subset of $\mathcal{L}(A, \mu)$. The aim of this section is to use the data $(X_i, Y_i)_{1 \leq i \leq n}$ in order to select a function $\hat{s} \in S$ close to the unknown conditional density s. We begin by presenting the procedure. The underlying motivations will be further discussed below.

Let $\bar{\Delta}$ be a map on S satisfying

$$\forall f \in S,\ \bar{\Delta}(f) \geq 1 \quad \text{and} \quad \sum_{f \in S} e^{-\bar{\Delta}(f)} \leq 1.$$

We define the function T on S^2 by

$$\begin{aligned} T(f,f') &= \frac{1}{n}\sum_{i=1}^{n}\frac{\sqrt{f'(X_i,Y_i)}-\sqrt{f(X_i,Y_i)}}{\sqrt{f(X_i,Y_i)+f'(X_i,Y_i)}} \\ &\quad +\frac{1}{2n}\sum_{i=1}^{n}\int_{A_2}\sqrt{f(X_i,y)+f'(X_i,y)}\left(\sqrt{f'(X_i,y)}-\sqrt{f(X_i,y)}\right)d\mu(y) \\ &\quad +\frac{1}{\sqrt{2}n}\sum_{i=1}^{n}\int_{A_2}\big(f(X_i,y)-f'(X_i,y)\big)\,d\mu(y), \end{aligned}$$

where the convention $0/0=0$ is used. We set for $L>0$,

$$\gamma(f)=\sup_{f'\in S}\left\{T(f,f')-L\frac{\bar{\Delta}(f')}{n}\right\}.$$

We finally define our estimator $\hat{s}\in S$ as any element of S such that

$$\gamma(\hat{s})+L\frac{\bar{\Delta}(\hat{s})}{n}\leq\inf_{f\in S}\left\{\gamma(f)+L\frac{\bar{\Delta}(f)}{n}\right\}+\frac{1}{n}. \tag{1}$$

Remarks. The definition of T comes from a decomposition of the Hellinger distance initiated by Baraud (2011) and taken back in Baraud (2013); Sart (2014, 2015); Baraud et al. (2016). We shall show in the proof of Theorem 1 that for all $f,f'\in\mathcal{L}(A,\mu)$, $\xi>0$, the two following assertions hold true with probability larger than $1-e^{-n\xi}$:

- If $T(f,f')\geq 0$, then $h^2(s,f')\leq c_1h^2(s,f)+c_2\xi$
- If $T(f,f')\leq 0$, then $h^2(s,f)\leq c_1h^2(s,f')+c_2\xi$.

In the above inequalities, c_1 and c_2 are positive universal constants. The sign of $T(f,f')$ allows thus to know which function among f and f' is the closest to s (up to the multiplicative constant c_1 and the remainder term $c_2\xi$). Note that comparing directly $h^2(s,f)$ to $h^2(s,f')$ is not straightforward in practice since s and h are both unknown to the statistician.

The definition of the criterion γ looks like the one proposed in Section 4.1 of Sart (2014) for estimating the transition density of a Markov chain as well as the one proposed in Baraud et al. (2016) for estimating one or several densities. The underlying idea is that $\gamma(f)+L\bar{\Delta}(f)/n$ is roughly between $h^2(s,f)$ and $h^2(s,f)+L\bar{\Delta}(f)/n$. It is thus natural to minimize $\gamma(\cdot)+L\bar{\Delta}(\cdot)/n$ to define an estimator $\hat{s}$ of s. To be more precise, when L is large enough, the proof of Theorem 1 shows that for all $\xi>0$, the following chained inequalities hold true with probability larger than $1-e^{-n\xi}$ uniformly for $f\in S$,

$$(1-\varepsilon)h^2(s,f)-R_1(\xi)\leq\gamma(f)+L\frac{\bar{\Delta}(f)}{n}\leq(1+\varepsilon)h^2(s,f)+2L\frac{\bar{\Delta}(f)}{n}+R_2(\xi)$$

where

$$\begin{aligned} R_1(\xi) &= \inf_{f'\in S}\left\{(1+\varepsilon)h^2(s,f')+L\frac{\bar{\Delta}(f')}{n}\right\}+c_3\xi \\ R_2(\xi) &= -(1-\varepsilon)h^2(s,S)+c_4\xi \end{aligned}$$

for universal constants $c_3 > 0$, $c_4 > 0$, $\varepsilon \in (0,1)$. We recall that $h^2(s,S)$ is the square of the Hellinger distance between the conditional density s and the set S, $h^2(s,S) = \inf_{f\in S} h^2(s,f)$. Therefore, as $\hat{s}$ satisfies (1),

$$\begin{aligned}(1-\varepsilon)h^2(s,\hat{s}) &\leq \gamma(\hat{s}) + L\frac{\bar{\Delta}(\hat{s})}{n} + R_1(\xi)\\ &\leq \inf_{f\in S}\left\{\gamma(f) + L\frac{\bar{\Delta}(f)}{n}\right\} + 1/n + R_1(\xi)\\ &\leq \inf_{f\in S}\left\{(1+\varepsilon)h^2(s,f) + 2L\frac{\bar{\Delta}(f)}{n}\right\} + 1/n + R_1(\xi) + R_2(\xi).\end{aligned}$$

Rewriting this last inequality and using that $\bar{\Delta} \geq 1$ yields:

Theorem 1. *There exists a universal constant L_0 such that if $L \geq L_0$, any estimator $\hat{s} \in S$ satisfying (1) satisfies*

$$\forall \xi > 0, \quad \mathbb{P}\left[h^2(s,\hat{s}) \leq C_1 \inf_{f\in S}\left\{h^2(s,f) + L\frac{\bar{\Delta}(f)}{n}\right\} + C_2\xi\right] \geq 1 - e^{-n\xi}, \tag{2}$$

where C_1, C_2 are universal positive constants. In particular,

$$\mathbb{E}\left[h^2(s,\hat{s})\right] \leq C_3 \inf_{f\in S}\left\{h^2(s,f) + L\frac{\bar{\Delta}(f)}{n}\right\},$$

where $C_3 > 0$ is universal.

Note that the marginal density f_X influences the performance of the estimator $\hat{s}$ through the Hellinger loss h only. Moreover, no information on f_X is needed to build the estimator.

We can interpret the condition $\sum_{f\in S} e^{-\bar{\Delta}(f)} \leq 1$ as a (sub)-probability on S. The more complex S, the larger the weights $\bar{\Delta}(f)$. When S is finite, one can choose $\bar{\Delta}(f) = |\log S|$, and the above inequality becomes

$$\mathbb{P}\left[h^2(s,\hat{s}) \leq C_1\left(h^2(s,S) + L\frac{|\log S|}{n}\right) + C_2\xi\right] \geq 1 - e^{-n\xi}.$$

The Hellinger quadratic risk of the estimator $\hat{s}$ can therefore be bounded from above by a sum of two terms (up to a multiplicative constant): the first one stands for the bias term while the second one stands for the estimation term.

Let us mention that assuming that S is a subset of $\mathcal{L}(A,\mu)$ is not restrictive. Indeed, if f belongs to $\mathbb{L}^1_+(A,\nu\otimes\mu)\setminus\mathcal{L}(A,\mu)$, we can set

$$\pi(f)(x,y) = \begin{cases}\frac{f(x,y)}{\int_{A_2} f(x,t)\,\mathrm{d}\mu(t)} & \text{if } \int_{A_2} f(x,t)\,\mathrm{d}\mu(t) > 1 \text{ and } \int_{A_2} f(x,t)\,\mathrm{d}\mu(t) < \infty\\ f(x,y) & \text{if } \int_{A_2} f(x,t)\,\mathrm{d}\mu(t) \leq 1\\ 0 & \text{if } \int_{A_2} f(x,t)\,\mathrm{d}\mu(t) = \infty.\end{cases}$$

The function $\pi(f)$ belongs to $\mathcal{L}(A,\mu)$ and does always better than f:

Proposition 2. *For all $f \in \mathbb{L}^1_+(A,\nu\otimes\mu)$,*

$$h^2(s,\pi(f)) \leq h^2(s,f).$$

Thereby, if S is only assumed to be a subset of $\mathbb{L}^1_+(A, \nu \otimes \mu)$, the procedure applies with $S' = \{\pi(f),\, f \in S\} \subset \mathcal{L}(A, \mu)$ in place of S (and with $\bar{\Delta}(\pi(f)) = \bar{\Delta}(f)$). The resulting estimator $\hat{s} \in S'$ then satisfies (2).

Remark: the procedure does not depend on the dominating measure ν. However, the set S, which must be chosen by the statistician, must satisfy the above assumption $S \subset \mathbb{L}^1_+(A, \nu \otimes \mu)$, which usually requires the knowledge of ν. Actually, this assumption can be slightly strengthened to deal with an unknown, but finite measure ν. This may be of interest when ν is the (unknown) marginal distribution of X_i (in which case $f_X = 1$). More precisely, let $\mathbb{L}^1_{+,sup}(A, \mu)$ be the set of non-negative measurable functions vanishing outside A such that

$$\sup_{x \in A_1} \int_{A_2} f(x, y)\, \mathrm{d}\mu(y) < \infty.$$

The assumption $S \subset \mathbb{L}^1_{+,sup}(A, \mu)$ can be satisfied without knowing ν and implies $S \subset \mathbb{L}^1_+(A, \nu \otimes \mu)$.

2.2. Hold-out. As a first application of our oracle inequality, we consider the situation in which the set S is a family of estimators built on a preliminary sample. We suppose therefore that we have at hand two independent samples of $Z = (X, Y)$: $\mathbf{Z}_1 = (Z_1, \dots, Z_n)$ and $\mathbf{Z}_2 = (Z_{n+1}, \dots, Z_{2n})$. This is equivalent to splitting an initial sample $(Z_1, \dots, Z_{2n})$ of size $2n$ into two equal parts: $\mathbf{Z}_1$ and $\mathbf{Z}_2$.

Let $\hat{S} = \{\hat{s}_\lambda,\, \lambda \in \Lambda\} \subset \mathbb{L}^1_+(A, \nu \otimes \mu)$ be an at most countable collection of estimators based only on the first sample $\mathbf{Z}_1$. In view of Proposition 2, we may assume, without loss of generality, that for all $\lambda \in \Lambda$,

$$\forall x \in A_1,\ \int_{A_2} \hat{s}_\lambda(x, y)\, \mathrm{d}\mu(y) \leq 1.$$

Let $\Delta \geq 1$ be a map defined on Λ such that $\sum_{\lambda \in \Lambda} e^{-\Delta(\lambda)} \leq 1$.

Conditionally to $\mathbf{Z}_1$, $\hat{S}$ is a deterministic set. We can therefore apply our selection rule to $S = \hat{S}$, $\bar{\Delta}(\hat{s}_\lambda) = \Delta(\lambda)$ and to the sample $\mathbf{Z}_2$ to derive an estimator $\hat{s}$ such that:

$$\forall \xi > 0, \quad \mathbb{P}\left[h^2(s, \hat{s}) \leq C_1 \inf_{\lambda \in \Lambda} \left\{h^2(s, \hat{s}_\lambda) + L\frac{\Delta(\lambda)}{n}\right\} + C_2\xi \;\middle|\; \mathbf{Z}_1\right] \geq 1 - e^{-n\xi}.$$

By taking the expectation with respect to $\mathbf{Z}_1$, we then deduce:

$$\forall \xi > 0, \quad \mathbb{P}\left[h^2(s, \hat{s}) \leq C_1 \inf_{\lambda \in \Lambda} \left\{h^2(s, \hat{s}_\lambda) + L\frac{\Delta(\lambda)}{n}\right\} + C_2\xi\right] \geq 1 - e^{-n\xi}. \tag{3}$$

Note that there is almost no assumption on the preliminary estimators. It is only assumed that $\hat{s}_\lambda \in \mathbb{L}^1_+(A, \nu \otimes \mu)$. Besides, the non-negativity of $\hat{s}_\lambda$ can always be fixed by taking its positive part if needed. We may therefore select among Kernel estimators (to choose the bandwith for instance), local polynomial estimators, projection estimators... It is also possible to mix in the collection $\{\hat{s}_\lambda,\, \lambda \in \Lambda\}$ several type of estimators. From a numerical point of view, the procedure can be implemented in practice provided that $|\Lambda|$ is finite and not too large.

We shall illustrate this result by applying it to some families of piecewise constant estimators. As we shall see, the resulting estimator $\hat{s}$ will be optimal and adaptive over some range of possibly anisotropic Hölder and possibly inhomogeneous Besov classes.

2.3. Histogram type estimators. We now define the piecewise constant estimators. Let m be a (finite) partition of $A \subset \mathbb{X} \times \mathbb{Y}$, and

$$\hat{s}_m(x,y) = \sum_{K \in m} \frac{\sum_{i=1}^n 1_K(X_i, Y_i)}{\sum_{i=1}^n (\delta_{X_i} \otimes \mu)(K)} 1_K(x,y),$$

where the conventions $0/0 = 0$, $x/\infty = 0$ are used. Györfi and Kohler (2007) established an integrated $\mathbb{L}^1$ risk bound for $\hat{s}_m$ under Lipschitz conditions on s. We are nevertheless unable to find in the literature a non-asymptotic risk bound for the Hellinger deterministic loss h. We propose the following result (which is assumption free on s):

Proposition 3. *Let m be a (finite) partition of A such that each $K \in m$ is of the form $I \times J$ with $I \subset A_1$, $J \subset A_2$ and $\mu(J) < \infty$. Let V_m be the cone of non-negative piecewise constant functions on the partition m defined by*

$$V_m = \left\{ \sum_{K \in m} a_K 1_K,\ \forall K \in m,\ a_K \in [0, +\infty) \right\}.$$

Then,

$$\mathbb{E}\left[h^2(s, \hat{s}_m)\right] \leq 4h^2(s, V_m) + 4\frac{|m|}{n}.$$

This result shows that the Hellinger quadratic risk $h^2(s, \hat{s}_m)$ of the estimator $\hat{s}_m$ can be bounded by a sum of two terms. The first one $h^2(s, V_m)$ corresponds to a bias term whereas the second one $|m|/n$ corresponds to a variance or estimation term. A deviation bound can also be established for some partitions:

Proposition 4. *Assume that m is a (finite) partition of A of the form*

$$m = \{I \times J,\ I \in \mathcal{I},\ J \in \mathcal{J}_I\},$$

where $\mathcal{I}$ is a (finite) partition of A_1, and, for each $I \in \mathcal{I}$, $\mathcal{J}_I$ is a (finite) partition of A_2 such that $\mu(J) < \infty$ for all $J \in \mathcal{J}_I$.

Then, there exist universal constants $C_1, C_2 > 0$ such that for all $\xi > 0$,

$$\mathbb{P}\left[h^2(s, \hat{s}_m) \leq 4h^2(s, V_m) + C_1\frac{|m|}{n} + C_2\xi\right] \geq 1 - e^{-n\xi}.$$

2.4. Selecting among piecewise constant estimators by Hold-out. The risk of a histogram type estimator $\hat{s}_m$ depends on the choice of the partition m: the thinner m, the smaller the bias term $h^2(s, V_m)$ but the larger the variance term $|m|/n$. Choosing a good partition m, that is a partition that realizes a good trade-off between the bias and variance terms is difficult in practice since $h^2(s, V_m)$ is unknown (as it involves the unknown conditional density s and the unknown distance h). Nevertheless, combining (3) and Proposition 3 immediately entails the following corollary.

Corollary 1. *Let $\mathcal{M}$ be an at most countable collection of finite partitions m of A. Assume that each $K \in m$ is of the form $I \times J$ with $I \subset A_1$, $J \subset A_2$ and $\mu(J) < \infty$. Let $\Delta \geq 1$ be a map on $\mathcal{M}$ satisfying*

$$\sum_{m \in \mathcal{M}} e^{-\Delta(m)} \leq 1.$$

Then, there exists an estimator $\hat{s}$ such that

$$\mathbb{E}\left[h^2\left(s,\hat{s}\right)\right] \leq C \inf_{m\in\mathcal{M}} \left\{ h^2\left(s,V_m\right) + \frac{|m|+\Delta(m)}{n} \right\}, \tag{4}$$

where C is a universal positive constant.

The novelty of this oracle inequality lies in the fact that it holds for an (unknown) deterministic Hellinger loss under very mild assumptions both on the partitions and the statistical setting. We avoid some classical assumptions that are required in the literature to prove similar inequalities (see, for instance, Theorem 3.1 of Akakpo and Lacour (2011) for a result with respect to a $\mathbb{L}^2$ loss).

2.5. Minimax rates over Hölder and Besov spaces. We can now deduce from (4) estimators with nice statistical properties under smoothness assumptions on the conditional density. Throughout this section, $\mathbb{X}\times\mathbb{Y}=\mathbb{R}^d$, $A=[0,1]^d$ and μ is the Lebesgue measure.

2.5.1. Hölder spaces. Given $\alpha\in(0,1]$, we recall that the Hölder space $\mathcal{H}^\alpha([0,1])$ is the set of functions f on $[0,1]$ for which there exists $|f|_\alpha>0$ such that

$$|f(x)-f(y)| \leq |f|_\alpha |x-y|^\alpha \quad \text{for all } x,y\in[0,1]. \tag{5}$$

Given $\boldsymbol{\alpha}=(\alpha_1,\dots,\alpha_d)\in(0,1]^d$, the Hölder space $\mathcal{H}^{\boldsymbol{\alpha}}([0,1]^d)$ is the set of functions f on $[0,1]^d$ such that for all $(x_1,\dots,x_d)\in(0,1]^d$, $j\in\{1,\dots,d\}$,

$$f_j(\cdot)=f(x_1,\cdots,x_{j-1},\cdot,x_{j+1},\dots x_d)$$

satisfies (5) with some constant $|f_j|_{\alpha_j}$ independent of $x_1,\dots,x_{j-1},x_{j+1},\dots,x_d$. We then set $|f|_{\boldsymbol{\alpha}}=\max_{1\leq j\leq d}|f_j|_{\alpha_j}$. When all the α_j are equals, the Hölder space $\mathcal{H}^{\boldsymbol{\alpha}}([0,1]^d)$ is said to be isotropic and anisotropic otherwise.

Choosing suitably the collection $\mathcal{M}$ of partitions allows to bound from above the right-hand side of (4) when $\sqrt{s}|_{[0,1]^d}$ is Hölderian. More precisely, for each integer $N\in\mathbb{N}^\star$, let m_N be the regular partition of $[0,1]$ with N pieces

$$m_N=\{[0,1/N[\,,[1/N,2/N[\,,\dots,[(N-1)/N,1]\}\,.$$

We may define for each multi-integer $\mathbf{N}=(N_1,\dots,N_d)\in(\mathbb{N}^\star)^d$,

$$m_{\mathbf{N}}=\left\{\prod_{j=1}^d I_j,\quad \forall j\in\{1,\dots,d\},\ I_j\in m_{N_j}\right\}.$$

We now choose $\mathcal{M}=\left\{m_{\mathbf{N}},\,\mathbf{N}\in(\mathbb{N}^\star)^d\right\}$, $\Delta(m_{\mathbf{N}})=|m_{\mathbf{N}}|$ to deduce (see, for instance, Lemma 4 and Corollary 2 of Birgé (2007) among numerous other references):

Corollary 2. *There exists an estimator $\hat{s}$ such that for all $\boldsymbol{\alpha}\in(0,1]^d$ and $\sqrt{s}|_{[0,1]^d}\in\mathcal{H}^{\boldsymbol{\alpha}}([0,1]^d)$,*

$$\mathbb{E}\left[h^2\left(s,\hat{s}\right)\right] \leq C\left[\left|\sqrt{s}|_{[0,1]^d}\right|_{\boldsymbol{\alpha}}^{\frac{2d}{d+2\bar{\boldsymbol{\alpha}}}} n^{-\frac{2\bar{\boldsymbol{\alpha}}}{2\bar{\boldsymbol{\alpha}}+d}} + n^{-1}\right],$$

where $\bar{\boldsymbol{\alpha}}$ stands for the harmonic mean of $\boldsymbol{\alpha}$

$$\frac{1}{\bar{\boldsymbol{\alpha}}} = \frac{1}{d}\sum_{i=1}^{d}\frac{1}{\alpha_i},$$

and where C is a positive constant depending only on d.

The estimator $\hat{s}$ achieves therefore the optimal rate of convergence over the anisotropic Hölder classes $\mathcal{H}^{\boldsymbol{\alpha}}([0,1]^d)$, $\boldsymbol{\alpha} \in (0,1]^d$. It is moreover adaptive since its construction does not involve the smoothness parameter $\boldsymbol{\alpha}$.

2.5.2. Besov spaces. The preceding result may be generalized to the Besov classes under a mild assumption on the design density.

We refer to Section 2.3 of Akakpo (2012) for a precise definition of the Besov spaces. According to the notations developed in this paper, $\mathcal{B}_q^{\boldsymbol{\alpha}}(\mathbb{L}^p([0,1]^d))$ stands for the Besov space with parameters $p>0$, $q>0$, and smoothness index $\boldsymbol{\alpha} \in (\mathbb{R}_+^\star)^d$. We denote its semi norm by $|\cdot|_{\boldsymbol{\alpha},p,q}$. This space is said to be homogeneous when $p \geq 2$ and inhomogeneous otherwise. It is said to be isotropic when all the α_j are equals and anisotropic otherwise. We now set for $p \in (0,+\infty]$,

$$\mathscr{B}^{\boldsymbol{\alpha}}(\mathbb{L}^p([0,1]^d)) = \begin{cases} \mathscr{B}_\infty^{\boldsymbol{\alpha}}(\mathbb{L}^p([0,1]^d)) & \text{if } p \in (0,1] \\ \mathscr{B}_p^{\boldsymbol{\alpha}}(\mathbb{L}^p([0,1]^d)) & \text{if } p \in (1,2) \\ \mathscr{B}_\infty^{\boldsymbol{\alpha}}(\mathbb{L}^p([0,1]^d)) & \text{if } p \in [2,+\infty) \\ \mathcal{H}^{\boldsymbol{\alpha}}([0,1]^d) & \text{if } p = \infty \end{cases}$$

and denote by $|\cdot|_{\boldsymbol{\alpha},p}$ the semi norm associated to the space $\mathscr{B}^{\boldsymbol{\alpha}}(\mathbb{L}^p([0,1]^d))$.

The algorithm of Akakpo (2012) provides a collection $\mathcal{M}$ of partitions m that allows to bound the right-hand side of (4) from above when $\sqrt{s}\big|_{[0,1]^d}$ belongs to a Besov space. More precisely:

Corollary 3. *Suppose that the (possibly unknown) density f_X of X_i is upper bounded by a (possibly unknown) constant κ and that ν is the Lebesgue measure.*

Then, there exists an estimator $\hat{s}$ such that, for all $p \in (2d/(d+2),+\infty]$, $\boldsymbol{\alpha} \in (0,1)^d$, $\bar{\boldsymbol{\alpha}} > d(1/p-1/2)_+$ and $\sqrt{s}\big|_{[0,1]^d} \in \mathscr{B}^{\boldsymbol{\alpha}}(\mathbb{L}^p([0,1]^d))$,

$$\mathbb{E}\left[h^2\left(s,\hat{s}\right)\right] \leq C\left[\left|\sqrt{s}\big|_{[0,1]^d}\right|_{\boldsymbol{\alpha},p}^{\frac{2d}{d+2\bar{\boldsymbol{\alpha}}}} n^{-\frac{2\bar{\boldsymbol{\alpha}}}{2\bar{\boldsymbol{\alpha}}+d}} + n^{-1}\right], \tag{6}$$

where $C>0$ depends only on κ,d,$\boldsymbol{\alpha}$,p and where $\bar{\boldsymbol{\alpha}}$ denotes the harmonic mean of $\boldsymbol{\alpha}$.

Remark: the control of the bias term $h(s,V_m)$ in (4) naturally involves a smoothness assumption on the square root of s instead of s. However, the regularity of the square root of s may be deduced from that of s. Indeed, we can prove that if $s \in \mathscr{B}_q^{\boldsymbol{\alpha}}(\mathbb{L}^p([0,1]^d))$ with $\boldsymbol{\alpha} \in (0,1)^d$ then $\sqrt{s} \in \mathscr{B}_{2q}^{\boldsymbol{\alpha}/2}(\mathbb{L}^{2p}([0,1]^d))$ and $|\sqrt{s}|_{\boldsymbol{\alpha}/2,2p,2q} \leq \sqrt{|s|_{\boldsymbol{\alpha},p,q}}$. If, additionally, s is positive on $[0,1]^d$, then $\sqrt{s}$ also belongs to $\mathscr{B}_q^{\boldsymbol{\alpha}}(\mathbb{L}^p([0,1]^d))$ and

$$|\sqrt{s}|_{\boldsymbol{\alpha},p,q} \leq \frac{|s|_{\boldsymbol{\alpha},p,q}}{2\sqrt{\inf_{x\in[0,1]^d} s(x)}}.$$

Under the assumption of Corollary 3, we deduce that if $s \in \mathscr{B}^{\boldsymbol{\alpha}}_{\infty}(\mathbb{L}^p([0,1]^d))$ for some $p \in (2d/(d+2), +\infty]$, $\boldsymbol{\alpha} \in (0,1)^d$, $\bar{\boldsymbol{\alpha}} > d(1/p - 1/2)_+$,

$$\mathbb{E}\left[h^2(s,\hat{s})\right] \leq C \min\left\{ \left(\frac{\left| s_{|[0,1]^d} \right|_{\boldsymbol{\alpha},p,\infty}}{\left(\inf_{x\in[0,1]^d} s(x) \right)^{1/2}} \right)^{\frac{2d}{d+2\bar{\boldsymbol{\alpha}}}} n^{-\frac{2\bar{\boldsymbol{\alpha}}}{2\bar{\boldsymbol{\alpha}}+d}}, \left| s_{|[0,1]^d} \right|^{\frac{d}{d+\bar{\boldsymbol{\alpha}}}}_{\boldsymbol{\alpha},p,\infty} n^{-\frac{\bar{\boldsymbol{\alpha}}}{\bar{\boldsymbol{\alpha}}+d}}, n^{-1} \right\},$$

where $C > 0$ depends only on κ,d,$\boldsymbol{\alpha}$,p.

3. Model selection

The construction of adaptive and optimal estimators over Hölder and Besov classes follows from the oracle inequality (4). This inequality is itself deduced from Theorem 1. Actually, this latter theorem can be applied in a different way to deduce a more general oracle inequality. We can then derive adaptive and (nearly) optimal estimators over more general classes of functions.

3.1. A general model selection theorem.

From now on, the following assumption holds.

Assumption 1. *The (possibly unknown) density f_X of X_i is bounded above by a (possibly unknown) constant κ. Moreover, $\nu(A_1) \leq 1$.*

Let $\mathbb{L}^2(A, \nu\otimes\mu)$ be the space of square integrable functions on A with respect to the product measure $\nu\otimes\mu$ endowed with the distance

$$d_2^2(f,f') = \int_A \left(f(x,y) - f'(x,y)\right)^2 \mathrm{d}\nu(x)\,\mathrm{d}\mu(y) \quad \text{for all } f, f' \in \mathbb{L}^2(A,\nu\otimes\mu).$$

We say that a subset V of $\mathbb{L}^2(A,\nu\otimes\mu)$ is a model if it is a finite dimensional linear space.

The discretization trick described in Section 4.2 of Sart (2014) can be adapted to our statistical setting. It leads to the theorem below.

Theorem 5. *Suppose that Assumption 1 holds. Let $\mathbb{V}$ be an at most countable collection of models. Let $\Delta \geq 1$ be a map on $\mathbb{V}$ satisfying*

$$\sum_{V\in\mathbb{V}} e^{-\Delta(V)} \leq 1.$$

Then, there exists an estimator $\hat{s}$ such that for all $\xi > 0$

$$(7)\quad \mathbb{P}\left[h^2(s,\hat{s}) \leq C\left(\inf_{V\in\mathbb{V}} \left\{ \kappa d_2^2\left(\sqrt{s}, V\right) + \frac{\Delta(V) + \dim(V)\log n}{n} \right\} + \frac{\kappa}{n^2} + \xi \right) \right] \geq 1 - e^{-n\xi},$$

where $C > 0$ is universal. In particular,

$$\mathbb{E}\left[h^2(s,\hat{s})\right] \leq C' \inf_{V\in\mathbb{V}} \left\{ d_2^2\left(\sqrt{s}, V\right) + \frac{\Delta(V) + \dim(V)\log n}{n} \right\},$$

where $C' > 0$ depends only on κ.

As in Theorem 1, the condition $\sum_{V\in\mathbb{V}} e^{-\Delta(V)} \leq 1$ has a Bayesian flavour since it can be interpreted as a (sub)-probability on $\mathbb{V}$. When $\mathbb{V}$ does not contain too many models per dimension, we can set $\Delta(V) = (\dim V)\log n$, in which case (7) becomes

$$\mathbb{P}\left[h^2(s,\hat{s}) \leq C''\left(\inf_{V\in\mathbb{V}}\left\{\kappa d_2^2\left(\sqrt{s},V\right) + \frac{\dim(V)\log n}{n}\right\} + \frac{\kappa}{n^2} + \xi\right)\right] \geq 1 - e^{-n\xi},$$

where C'' is universal.

This theorem is more general than Corollary 1 since it enables us to deal with more general models V. Moreover, it provides a deviation bound for $h^2(s,\hat{s})$, which is not the case of Corollary 1. As a counterpart, it requires an assumption on the marginal density f_X and the bound involves a logarithmic term and κ.

Another difference between this theorem and Corollary 1 lies in the computation time of the estimators. The estimator of Corollary 1 may be built in practice in a reasonable amount of time if $|\mathcal{M}|$ is not too large. On the opposite, the procedure leading to the above estimator (which is described in the proof of the theorem) is numerically very expensive, and it is unlikely that it could be implemented in a reasonable amount of time. This estimator should therefore be only considered for theoretical purposes.

3.2. From model selection to estimation. It is recognized that a model selection theorem such as Theorem 5 is a bridge between statistics and approximation theory. Indeed, it remains to choose models with good approximation properties with respect to the assumptions we wish to consider on s to automatically derive a good estimator $\hat{s}$.

A convenient way to model these assumptions is to consider a class $\mathscr{F}$ of functions of $\mathbb{L}^2(A, \nu\otimes\mu)$ and to suppose that $\sqrt{s}|_A$ belongs to $\mathscr{F}$. The aim is then to choose $(\mathbb{V},\Delta)$ and to bound

$$\varepsilon_{\mathscr{F}}(f) = \inf_{V\in\mathbb{V}}\left\{d_2^2\left(f,V\right) + \frac{\Delta(V) + \dim(V)\log n}{n}\right\} \quad \text{for all } f\in\mathscr{F}$$

from above since

$$\begin{aligned}\mathbb{E}\left[h^2(s,\hat{s})\right] &\leq C'\varepsilon_{\mathscr{F}}(\sqrt{s})\\ \mathbb{P}\left[h^2(s,\hat{s}) \leq C''\varepsilon_{\mathscr{F}}(\sqrt{s}) + C'''\xi\right] &\geq 1 - e^{-n\xi} \quad \text{for all } \xi > 0\end{aligned}$$

where C', C'' depend only on κ and where C''' is universal. This work has already been carried out in the literature for different classes $\mathscr{F}$ of interest. The flexibility of our approach enables the study of various assumptions as illustrated by the three examples below. We refer to Sart (2014); Baraud and Birgé (2014) for additional examples. In the remainder of this section, μ and ν stand for the Lebesgue measure.

Besov classes. We suppose that $\mathbb{X}\times\mathbb{Y} = \mathbb{R}^d$, $A = [0,1]^d$ and that $\mathscr{F}$ is the class of smooth functions defined by

$$\mathscr{F} = \mathscr{B}([0,1]^d) = \bigcup_{p\in(0,+\infty)}\left(\bigcup_{\substack{\boldsymbol{\alpha}\in(0,+\infty)^d\\ \bar{\boldsymbol{\alpha}}>d(1/p-1/2)_+}} \mathscr{B}^{\boldsymbol{\alpha}}(\mathbb{L}^p([0,1]^d))\right).$$

It is then shown in Sart (2014) that one can choose a collection $\mathbb{V}$ provided by Theorem 1 of Akakpo (2012) to get:

$$\text{for all } f \in \mathscr{B}([0,1]^d), \quad \varepsilon_{\mathscr{F}}(f) \leq C \left[|f|_{\boldsymbol{\alpha},p}^{2d/(d+2\bar{\boldsymbol{\alpha}})} \left(\frac{\log n}{n} \right)^{2\bar{\boldsymbol{\alpha}}/(2\bar{\boldsymbol{\alpha}}+d)} + \frac{\log n}{n} \right], \tag{8}$$

where $p \in (0,+\infty)$, $\boldsymbol{\alpha} \in (0,+\infty)^d$, $\bar{\boldsymbol{\alpha}} > d(1/p-1/2)_+$ are such that $f \in \mathscr{B}^{\boldsymbol{\alpha}}(\mathbb{L}^p([0,1]^d))$ and where $C>0$ depends only on d, p, $\boldsymbol{\alpha}$.

With this choice of models, the estimator $\hat{s}$ of Theorem 5 converges at the expected rate (up to a logarithmic term) for the Hellinger deterministic loss h over a very wide range of possibly inhomogeneous and anisotropic Besov spaces. It is moreover adaptive with respect to the (possibly unknown) regularity index $\boldsymbol{\alpha}$ of $\sqrt{s}|_{[0,1]^d}$.

Regression model. We can also tackle the celebrated regression model $Y_i = g(X_i)+\varepsilon_i$ where g is an unknown function and where ε_i is an unobserved random variable. For the sake of simplicity, $\mathbb{X} = \mathbb{Y} = \mathbb{R}$, $A_1 = A_2 = [0,1]$. The conditional density s is of the form $s(x,y) = \varphi(y-g(x))$ where φ is the density of ε_i with respect to the Lebesgue measure.

Since φ and g are unknown, we can, for instance, suppose that these functions are smooth, which amounts to saying that $\sqrt{s}|_{[0,1]^2}$ belongs to

$$\mathscr{F} = \bigcup_{\alpha>0} \left\{ f,\, \exists \phi \in \mathcal{H}^\alpha(\mathbb{R}), \exists g \in \mathscr{B}([0,1]), \|g\|_\infty < \infty,\, \forall x,y \in [0,1],\ f(x,y) = \phi(y-g(x)) \right\}.$$

Here, $\mathcal{H}^\alpha(\mathbb{R})$ stands for the space of Hölderian functions on $\mathbb{R}$ with regularity index $\alpha \in (0,+\infty)$ and semi norm $|\cdot|_{\alpha,\infty}$. The notation $\|\cdot\|_\infty$ stands for the supremum norm: $\|g\|_\infty = \sup_{x\in[0,1]} |g(x)|$. An upper bound for $\varepsilon_{\mathscr{F}}(f)$ may be found in Section 4.4 of Sart (2014). Actually, we show in Section 4.6 that this bound can be slightly improved. To be more precise, the result is the following: for all $\alpha>0$, $p \in (0,+\infty]$, $\beta > (1/p-1/2)_+$, $\phi \in \mathcal{H}^\alpha(\mathbb{R})$, $g \in \mathscr{B}^\beta(\mathbb{L}^p([0,1]))$, such that $\|g\|_\infty < \infty$, and all function $f \in \mathscr{F}$ of the form $f(x,y) = \phi(y-g(x))$,

$$\varepsilon_{\mathscr{F}}(f) \leq C_1 \left(\frac{\log n}{n} \right)^{\frac{2\beta(\alpha\wedge 1)}{2\beta(\alpha\wedge 1)+1}} + C_2 \left(\frac{\log n}{n} \right)^{\frac{2\alpha}{2\alpha+1}}, \tag{9}$$

where C_1 depends only on p, β, α, $|g|_{\beta,p}$, $\|g\|_\infty$, $|\phi|_{\alpha\wedge 1,\infty}$ and where C_2 depends only on α, $\|g\|_\infty$, $|\phi|_{\alpha,\infty}$.

In particular, if ϕ is more regular than g in the sense that $\alpha \geq \beta \vee 1$, then the rate for estimating the conditional density s is the same as the one for estimating the regression function g (up to a logarithmic term). As shown in Sart (2014), this rate is always faster than the rate we would obtain under smoothness assumptions only that would ignore the specific form of s.

Remark. The reader could find in Sart (2014) a bound for $\varepsilon_{\mathscr{F}}$ when $\mathscr{F}$ corresponds to the heteroscedastic regression model $Y_i = g_1(X_i) + g_2(X_i)\varepsilon_i$, where g_1, g_2 are smooth unknown functions.

A single index type model. In this last example, we investigate the situation in which the explanatory random variables X_i lie in a high dimensional linear space, say $\mathbb{X} = \mathbb{R}^{d_1}$ with d_1 large. On the contrary, the random variables Y_i lie in a small dimensional linear space, say $\mathbb{Y} = \mathbb{R}^{d_2}$ with d_2 small. Our aim is then to estimate s on $A = A_1 \times A_2 = [0,1]^{d_1} \times [0,1]^{d_2}$.

It is well known (and this appears in (8)) that the curse of dimensionality prevents us to get fast rate of convergence under pure smoothness assumptions on s. A solution to overcome this difficulty is to use a single index approach as proposed by Hall and Yao (2005); Fan et al. (2009), that is to suppose that the conditional distribution $\mathcal{L}(Y_i \mid X_i = x)$ depends on x through an unknown parameter $\theta \in \mathbb{R}^{d_1}$. More precisely, we suppose in this section that s is of the form $s(x,y) = \varphi(<\theta, x>, y)$ where $<\cdot,\cdot>$ denotes the usual scalar product on $\mathbb{R}^{d_1}$ and where φ is a smooth unknown function. Without loss of generality, we can suppose that θ belongs to the unit ℓ^1 ball of $\mathbb{R}^{d_1}$ denoted by $\mathcal{B}_1(0,1)$. We can reformulate these different assumptions by saying that $\sqrt{s}\big|_{[0,1]^{d_1+d_2}}$ belongs to the set

$$\begin{aligned}\mathscr{F} &= \bigcup_{\boldsymbol{\alpha}\in(0,+\infty)^{1+d_2}} \Big\{ f,\, \exists g \in \mathcal{H}^{\boldsymbol{\alpha}}([0,1]^{1+d_2}), \exists \theta \in \mathcal{B}_1(0,1), \\ &\qquad\qquad \forall (x,y) \in [0,1]^{d_1+d_2},\ f(x,y) = g(<\theta, x>, y) \Big\}.\end{aligned}$$

A collection of models V possessing nice approximation properties with respect to the elements f of $\mathscr{F}$ can be built by using the results of Baraud and Birgé (2014). We prove in Section 4.6 that we can bound $\varepsilon_{\mathscr{F}}(f)$ as follows: for all $\boldsymbol{\alpha} \in (0,+\infty)^{1+d_2}$, $g \in \mathcal{H}^{\boldsymbol{\alpha}}([0,1]^{1+d_2})$, $\theta \in \mathcal{B}_1(0,1)$, and all function $f \in \mathscr{F}$ of the form $f(x,y) = g(<\theta, x>, y)$,

$$\varepsilon_{\mathscr{F}}(f) \le C_1 \, |g|_{\boldsymbol{\alpha},\infty}^{\frac{2(1+d_2)}{1+d_2+2\bar{\boldsymbol{\alpha}}}} \left(\frac{\log n}{n}\right)^{\frac{2\bar{\boldsymbol{\alpha}}}{2\bar{\boldsymbol{\alpha}}+1+d_2}} + C_2 d_1 \frac{\log n \vee \log\left(|g|^2_{\alpha_1\wedge 1,\infty}/d_1\right)}{n}, \tag{10}$$

where C_1 depends only on d_2, $\boldsymbol{\alpha}$, and where C_2 depends only on d_2, α_1. Although s is a function of $d_1 + d_2$ variables, the rate of convergence of $\hat{s}$ corresponds to the estimation rate of a smooth function g of $1 + d_2$ variables only (up to a logarithmic term).

4. Proofs

4.1. Proof of Theorem 1.

Lemma 1. *For all $f, f' \in S$, and $\xi > 0$, there exists an event $\Omega_\xi(f,f')$ such that $\mathbb{P}\left[\Omega_\xi(f,f')\right] \ge 1 - e^{-n\xi}$ and on which:*

$$(1-\varepsilon)\, h^2\left(s, f'\right) + T\left(f, f'\right) \le (1+\varepsilon)\, h^2\left(s, f\right) + L_1 \xi,$$

where $L_1 > 0$, $\varepsilon \in (0,1)$ are positive universal constants.

Proof. Let ψ_1 and ψ_2 be the functions defined on $(\mathbb{R}_+)^2$ by

$$\begin{aligned}\psi_1(x,y) &= \frac{\sqrt{y} - \sqrt{x}}{\sqrt{x+y}} \\ \psi_2(x,y) &= \sqrt{\frac{x+y}{2}} - \left(\sqrt{x} + \sqrt{y}\right)\end{aligned}$$

where the convention $0/0 = 0$ is used. Let

$$\begin{aligned}T_{1,i}(f,f') &= \psi_1\left(f(X_i,Y_i), f'(X_i,Y_i)\right) \\ T_{2,i}(f,f') &= \frac{1}{\sqrt{2}} \int_{A_2} \psi_2\left(f(X_i,y), f'(X_i,y)\right)\left(\sqrt{f'(X_i,y)} - \sqrt{f(X_i,y)}\right) d\mu(y).\end{aligned}$$

We decompose $T(f,f')$ as

$$T(f,f') = \frac{1}{n}\sum_{i=1}^{n}\left(T_{1,i}(f,f') + T_{2,i}(f,f')\right)$$

and define $Z(f,f') = T(f,f') - \mathbb{E}[T(f,f')]$.

We need the claim below whose proof requires the same arguments than those developed in the proofs of Corollary 1 and Proposition 3 of Baraud (2011). As these arguments are short, we make them explicit at the end of this section to make the paper self contained.

Claim 1. *For all $f, f' \in S$,*

$$\left(\sqrt{2}-1\right)h^2\left(s,f'\right) + T\left(f,f'\right) \leq \left(1+\sqrt{2}\right)h^2\left(s,f\right) + Z\left(f,f'\right) \tag{11}$$

$$\mathbb{E}\left[T_{1,i}^2\left(f,f'\right)\right] \leq 6\left[h^2(s,f) + h^2(s,f')\right]. \tag{12}$$

By using Cauchy-Schwarz inequality,

$$\begin{aligned}\left(T_{2,i}(f,f')\right)^2 \leq{}& \left(\int_{A_2}\left(\psi_2\left(f(X_i,y), f'(X_i,y)\right)\right)^2 \mathrm{d}\mu(y)\right)\times \\ & \left(\frac{1}{2}\int_{A_2}\left(\sqrt{f'(X_i,y)} - \sqrt{f(X_i,y)}\right)^2 d\mu(y)\right).\end{aligned} \tag{13}$$

Note that the function

$$z \in [0,\infty) \mapsto \frac{\sqrt{1+z}}{1+\sqrt{z}}$$

is bounded below by $1/\sqrt{2}$ and bounded above by 1. Therefore, for all $z \geq 0$,

$$\begin{aligned}\frac{1}{\sqrt{2}} \leq \frac{\sqrt{1+z}}{1+\sqrt{z}} \leq 1 &\iff \frac{1+\sqrt{z}}{2} \leq \sqrt{\frac{1+z}{2}} \leq \frac{1+\sqrt{z}}{\sqrt{2}} \\ &\iff -\frac{1+\sqrt{z}}{2} \leq \sqrt{\frac{1+z}{2}} - (1+\sqrt{z}) \leq \frac{1-\sqrt{2}}{\sqrt{2}}\left(1+\sqrt{z}\right) \\ &\implies \left|\sqrt{\frac{1+z}{2}} - (1+\sqrt{z})\right| \leq \frac{1+\sqrt{z}}{2}.\end{aligned}$$

For all $x, y \geq 0$, we derive from this inequality with $z = x/y$ that

$$|\psi_2(x,y)| \leq \frac{\sqrt{x}+\sqrt{y}}{2} \quad \text{for all } x,y \geq 0.$$

Thereby,

$$(\psi_2(x,y))^2 \leq \frac{(\sqrt{x}+\sqrt{y})^2}{4} \leq \frac{x+y}{2},$$

which together with $f, f' \in \mathcal{L}(A,\mu)$ and (13) yields

$$\begin{aligned}\mathbb{E}\left[\left(T_{2,i}(f,f')\right)^2\right] &\leq h^2(f,f') \\ &\leq 2h^2(s,f) + 2h^2(s,f').\end{aligned}$$

By using (12), we get

$$\begin{aligned}\mathbb{E}\left[\left(T_{1,i}(f,f')+T_{2,i}(f,f')\right)^2\right] &\leq 2\mathbb{E}\left[\left(T_{1,i}(f,f')\right)^2+\left(T_{2,i}(f,f')\right)^2\right]\\ &\leq 16h^2(s,f)+16h^2(s,f').\end{aligned}$$

Now, $T_{1,i}(f,f')\leq 1$ as ψ_1 is bounded by 1 and

$$\begin{aligned}T_{2,i}(f,f') &\leq \frac{1}{\sqrt{2}}\int_{A_2}\left|\frac{\sqrt{f'(X_i,y)}+\sqrt{f'(X_i,y)}}{2}\right|\left|\sqrt{f'(X_i,y)}-\sqrt{f(X_i,y)}\right|d\mu(y)\\ &\leq \frac{1}{\sqrt{2}}\int_{A_2}\frac{f'(X_i,y)+f(X_i,y)}{2}d\mu(y)\\ &\leq \frac{1}{\sqrt{2}}.\end{aligned}$$

Bernstein's inequality and more precisely equation (2.20) of Massart (2003) shows that for all $\xi>0$,

$$\mathbb{P}\left[Z(f,f')\leq\sqrt{32\left(h^2(s,f)+h^2(s,f')\right)\xi}+(1+1/\sqrt{2})\xi/3\right]\geq 1-e^{-n\xi}.$$

Using now that

$$2\sqrt{xy}\leq\alpha x+\alpha^{-1}y \quad \text{for all } x,y\geq 0 \text{ and } \alpha>0,$$

we get with probability larger than $1-e^{-n\xi}$,

$$Z(f,f')\leq\alpha\sqrt{8}\left(h^2(s,f)+h^2(s,f')\right)+\left((1+1/\sqrt{2})/3+\sqrt{8}\alpha^{-1}\right)\xi.$$

Therefore, we deduce from (11),

$$\left(\sqrt{2}-1-\alpha\sqrt{8}\right)h^2\left(s,f'\right)+T\left(f,f'\right)\leq\left(\sqrt{2}+1+\alpha\sqrt{8}\right)h^2\left(s,f\right)+\left((1+1/\sqrt{2})/3+\sqrt{8}\alpha^{-1}\right)\xi.$$

It remains to choose α to complete the proof. Any value $\alpha\in(0,(\sqrt{2}-1)/\sqrt{8})$ works. □

Lemma 2. *For all $\xi>0$ and $f\in S$, there exists an event $\Omega_\xi(f)$ such that $\mathbb{P}\left[\Omega_\xi(f)\right]\geq 1-e^{-n\xi}$ and on which:*

$$\forall f'\in S,\quad (1-\varepsilon)\,h^2\left(s,f'\right)+T\left(f,f'\right)\leq(1+\varepsilon)\,h^2\left(s,f\right)+L_1\frac{\bar{\Delta}(f')}{n}+L_1\xi, \tag{14}$$

where $L_1>0$, $\varepsilon\in(0,1)$ are given in Lemma 1. Moreover, there exists an event Ω_ξ such that $\mathbb{P}\left[\Omega_\xi\right]\geq 1-e^{-n\xi}$ and on which:

$$\forall f,f'\in S,\ (1-\varepsilon)\,h^2\left(s,f'\right)+T\left(f,f'\right)\leq(1+\varepsilon)\,h^2\left(s,f\right)+L_1\frac{\bar{\Delta}(f')}{n}+L_1\frac{\bar{\Delta}(f)}{n}+L_1\xi. \tag{15}$$

Proof. The result follows easily from Lemma 1 by setting

$$\Omega_\xi(f)=\bigcap_{f'\in S}\Omega_{\xi+\frac{\bar{\Delta}(f')}{n}}(f,f') \quad\text{and}\quad \Omega_\xi=\bigcap_{f\in S}\Omega_{\xi+\frac{\bar{\Delta}(f)}{n}}(f).$$

□

Lemma 3. *Set $L_0 = (1+\log 2)L_1$ where L_1 is given in Lemma 1. For all $\xi > 0$, the following holds with probability larger than $1 - e^{-n\xi}$: if $L \geq L_0$, for all $f \in S$,*

$$(1-\varepsilon)h^2(s,f) - R_1(\xi) \leq \gamma(f) + L\frac{\bar{\Delta}(f)}{n} \leq (1+\varepsilon)h^2(s,f) + 2L\frac{\bar{\Delta}(f)}{n} + R_2(\xi) \tag{16}$$

where

$$\begin{aligned} R_1(\xi) &= \inf_{f' \in S}\left\{(1+\varepsilon)h^2(s,f') + L\frac{\bar{\Delta}(f')}{n}\right\} + c_1\xi \\ R_2(\xi) &= -(1-\varepsilon)h^2(s,S) + c_2\xi \end{aligned}$$

and where ε is given in Lemma 1, and c_1 and c_2 are universal positive constants ($c_1 = 2L_1$ and $c_2 = L_1$).

Proof. Let $g \in S$ be such that

$$(1+\varepsilon)h^2(s,g) + L\frac{\bar{\Delta}(g)}{n} \leq \inf_{f' \in S}\left\{(1+\varepsilon)h^2\left(s,f'\right) + L\frac{\bar{\Delta}(f')}{n}\right\} + L_1\xi.$$

We shall show that (16) holds on the event $\Omega_{\xi+\frac{\log 2}{n}}(g) \cap \Omega_{\xi+\frac{\log 2}{n}}$. We derive from (15) that for all $f, f' \in S$,

$$\begin{aligned} (1-\varepsilon)\,h^2\left(s,f'\right) + \left(T\left(f,f'\right) - L\frac{\bar{\Delta}(f')}{n}\right) &\leq (1+\varepsilon)\,h^2(s,f) + L_1\frac{\bar{\Delta}(f)}{n} + L_1\frac{\log 2}{n} + L_1\xi \\ &\leq (1+\varepsilon)\,h^2(s,f) + L\frac{\bar{\Delta}(f)}{n} + L_1\xi, \end{aligned}$$

which in particular implies

$$\begin{aligned} \gamma(f) &\leq (1+\varepsilon)\,h^2(s,f) + L\frac{\bar{\Delta}(f)}{n} - (1-\varepsilon)h^2(s,S) + L_1\xi \\ &\leq (1+\varepsilon)\,h^2(s,f) + L\frac{\bar{\Delta}(f)}{n} + R_2(\xi). \end{aligned}$$

This proves the right inequality of (16). We now turn to the left one. We use (14) to get for all $f \in S$,

$$\begin{aligned} (1-\varepsilon)\,h^2(s,f) &\leq (1+\varepsilon)\,h^2(s,g) + T(f,g) + L_1\frac{\bar{\Delta}(f)}{n} + L_1\frac{\log 2}{n} + L_1\xi \\ &\leq (1+\varepsilon)\,h^2(s,g) + L\frac{\bar{\Delta}(g)}{n} + \left(T(f,g) - L\frac{\bar{\Delta}(g)}{n}\right) + L\frac{\bar{\Delta}(f)}{n} + L_1\xi \\ &\leq (1+\varepsilon)\,h^2(s,g) + L\frac{\bar{\Delta}(g)}{n} + \gamma(f) + L\frac{\bar{\Delta}(f)}{n} + L_1\xi \\ &\leq \inf_{f' \in S}\left\{(1+\varepsilon)h^2\left(s,f'\right) + L\frac{\bar{\Delta}(f')}{n}\right\} + \gamma(f) + L\frac{\bar{\Delta}(f)}{n} + 2L_1\xi. \end{aligned}$$

This ends the proof. □

The computations preceding Theorem 1 finally complete its proof. □

Proof of Claim 1. Define the function $g = (f + f')/2$ and the measure ζ by

$$\mathrm{d}\zeta(x, y) = f_X(x)\,\mathrm{d}\nu(x)\,\mathrm{d}\mu(y).$$

We have,

$$\begin{aligned}
\mathbb{E}\left[T\left(f, f'\right)\right] &= \frac{1}{\sqrt{2}}\int_A \frac{\sqrt{f'} - \sqrt{f}}{\sqrt{g}} s\,\mathrm{d}\zeta + \frac{1}{\sqrt{2}}\int_A \sqrt{g}\left(\sqrt{f'} - \sqrt{f}\right)\,\mathrm{d}\zeta + \frac{1}{\sqrt{2}}\int_A \left(f - f'\right)\,\mathrm{d}\zeta \\
&= \frac{1}{\sqrt{2}}\left(\int_A \sqrt{\frac{f'}{g}} s\,\mathrm{d}\zeta + \int_A \sqrt{gf'}\,\mathrm{d}\zeta - \int_A f'\,\mathrm{d}\zeta\right) \\
&\quad -\frac{1}{\sqrt{2}}\left(\int_A \sqrt{\frac{f}{g}} s\,\mathrm{d}\zeta + \int_A \sqrt{gf}\,\mathrm{d}\zeta - \int_A f\,\mathrm{d}\zeta\right).
\end{aligned}$$

Now,

$$\begin{aligned}
h^2(s, f') - h^2(s, f) &= \left(\int_A \sqrt{sf}\,\mathrm{d}\zeta - \frac{1}{2}\int_A f\,\mathrm{d}\zeta\right) - \left(\int_A \sqrt{sf'}\,\mathrm{d}\zeta - \frac{1}{2}\int_A f'\,\mathrm{d}\zeta\right) \\
&= -\frac{1}{\sqrt{2}}\mathbb{E}\left[T(f, f')\right] + \frac{1}{2}\left(\int_A \sqrt{\frac{f'}{g}} s\,\mathrm{d}\zeta + \int_A \sqrt{gf'}\,\mathrm{d}\zeta - 2\int_A \sqrt{sf'}\,\mathrm{d}\zeta\right) \\
&\quad -\frac{1}{2}\left(\int_A \sqrt{\frac{f}{g}} s\,\mathrm{d}\zeta + \int_A \sqrt{gf}\,\mathrm{d}\zeta - 2\int_A \sqrt{sf}\,\mathrm{d}\zeta\right) \\
&= -\frac{1}{\sqrt{2}}\mathbb{E}\left[T(f, f')\right] + \frac{1}{2}\int_A \sqrt{\frac{f'}{g}}\left(\sqrt{s} - \sqrt{g}\right)^2\,\mathrm{d}\zeta - \frac{1}{2}\int_A \sqrt{\frac{f}{g}}\left(\sqrt{s} - \sqrt{g}\right)^2\,\mathrm{d}\zeta \\
&\leq -\frac{1}{\sqrt{2}}\mathbb{E}\left[T(f, f')\right] + \frac{1}{2}\int_A \sqrt{\frac{f'}{g}}\left(\sqrt{s} - \sqrt{g}\right)^2\,\mathrm{d}\zeta.
\end{aligned}$$

By using $\sqrt{f'/g} \leq \sqrt{2}$, and a concavity argument,

$$\begin{aligned}
\frac{1}{2}\int_A \sqrt{\frac{f'}{g}}\left(\sqrt{s} - \sqrt{g}\right)^2\,\mathrm{d}\zeta &\leq \sqrt{2}h^2(s, g) \\
&\leq \frac{1}{\sqrt{2}}\left(h^2(s, f) + h^2(s, f')\right).
\end{aligned}$$

We now derive from $-\mathbb{E}\left[T(f, f')\right] = -T(f, f') + Z(f, f')$ that,

$$h^2(s, f') - h^2(s, f) \leq \frac{-T(f, f') + Z(f, f')}{\sqrt{2}} + \frac{1}{\sqrt{2}}\left(h^2(s, f) + h^2(s, f')\right).$$

This proves (11).

We now turn to the proof of (12):

$$\begin{aligned}\mathbb{E}\left[T_{1,i}^2\left(f,f'\right)\right] &= \frac{1}{2}\int_A \frac{\left(\sqrt{f'}-\sqrt{f}\right)^2}{g} s\,\mathrm{d}\zeta \\ &= \frac{1}{2}\int_A \left(\sqrt{f'}-\sqrt{f}\right)^2\left(\frac{\sqrt{s}-\sqrt{g}}{\sqrt{g}}+1\right)^2\,\mathrm{d}\zeta \\ &\leq \int_A \frac{\left(\sqrt{f'}-\sqrt{f}\right)^2}{g}\left(\sqrt{s}-\sqrt{g}\right)^2\,\mathrm{d}\zeta + \int_A \left(\sqrt{f'}-\sqrt{f}\right)^2\,\mathrm{d}\zeta.\end{aligned}$$

Now, $(\sqrt{f'}-\sqrt{f})^2/g \leq 2$ and hence,

$$\begin{aligned}\mathbb{E}\left[T_{1,i}^2\left(f,f'\right)\right] &\leq 2\int_A \left(\sqrt{s}-\sqrt{g}\right)^2\,\mathrm{d}\zeta + \int_A \left(\sqrt{f'}-\sqrt{f}\right)^2\,\mathrm{d}\zeta \\ &\leq 4h^2(s,g)+2h^2(f,f') \\ &\leq 4h^2(s,g)+4h^2(s,f)+4h^2(s,f').\end{aligned}$$

By using a concavity argument, $h^2(s,g)\leq 1/2(h^2(s,f)+h^2(s,f'))$. Finally,

$$\mathbb{E}\left[T_{1,i}^2\left(f,f'\right)\right] \leq 6\left[h^2(s,f)+h^2(s,f')\right],$$

which proves (12). □

4.2. Proof of Proposition 2. As f belongs to $\mathbb{L}^1_+(A,\nu\otimes\mu)$, Fubini's theorem says that there exists $A_1'\subset A_1$ such that $\nu(A_1\setminus A_1')=0$ and such that

$$\forall x\in A_1',\quad \int_{A_2} f(x,y)\,\mathrm{d}\mu(y)<\infty.$$

Let $(\mathbb{L}^2(A_2,\mu),\|\cdot\|)$ be the linear space of square integrable functions on A_2 with respect to μ.

For all $x\in A_1'$, $\sqrt{f(x,\cdot)}$ belongs to $\mathbb{L}^2(A_2,\mu)$ and

$$\sqrt{\pi(f)(x,y)} = \frac{\sqrt{f(x,y)}}{\max\left(\|\sqrt{f(x,\cdot)}\|,1\right)} \quad \text{for all } (x,y)\in A_1'\times A_2.$$

Note that $\sqrt{\pi(f)(x,\cdot)}$ is the projection of $\sqrt{f(x,\cdot)}$ onto the unit ball $\{g\in\mathbb{L}^2(A_2,\mu),\,\|g\|\leq 1\}$. As the projection is Lipschitz continuous,

$$\left\|\sqrt{\pi(s)(x,\cdot)}-\sqrt{\pi(f)(x,\cdot)}\right\|^2 \leq \left\|\sqrt{s(x,\cdot)}-\sqrt{f(x,\cdot)}\right\|^2 \quad \text{for all } x\in A_1'.$$

As $\|\sqrt{s(x,\cdot)}\|\leq 1$, $\sqrt{\pi(s)(x,\cdot)}=\sqrt{s(x,\cdot)}$ and hence

$$\left\|\sqrt{s(x,\cdot)}-\sqrt{\pi(f)(x,\cdot)}\right\|^2 \leq \left\|\sqrt{s(x,\cdot)}-\sqrt{f(x,\cdot)}\right\|^2 \quad \text{for all } x\in A_1'.$$

By integrating both inequalities with respect to x,

$$\begin{aligned}\int_{A_1'}\int_{A_2}\left(\sqrt{s(x,\cdot)}-\sqrt{\pi(f)(x,y)}\right)^2\,\mathrm{d}\nu(x)\,\mathrm{d}\mu(y) &\leq \int_{A_1'}\int_{A_2}\left(\sqrt{s(x,\cdot)}-\sqrt{f(x,y)}\right)^2\,\mathrm{d}\nu(x)\,\mathrm{d}\mu(y) \\ &\leq 2h^2(s,f).\end{aligned}$$

Since $\nu(A_1\setminus A_1')=0$, the left-hand side of the above inequality is merely $2h^2(s,\pi(f))$, which proves the proposition. □

4.3. Proof of Proposition 3. Let for each $K \in m$, $I_K \subset A_1$ and $J_K \subset A_2$ be such that $K = I_K \times J_K$. Let $\mathcal{I} = \{I_K,\, K \in m\}$, and for each $I \in \mathcal{I}$, let $\mathcal{J}_I = \{J,\, I \times J \in m\}$. The partition m can be rewritten as

$$m = \bigcup_{I \in \mathcal{I}} \{I \times J,\, J \in \mathcal{J}_I\}\,.$$

Remark that

$$|m| = \sum_{I \in \mathcal{I}} |\mathcal{J}_I|.$$

We now introduce for all $I \in \mathcal{I}$ and $J \in \mathcal{J}_I$,

$$N(I \times J) = \sum_{i=1}^{n} 1_I(X_i) 1_J(Y_i) \quad \text{and} \quad M(I) = \sum_{i=1}^{n} 1_I(X_i).$$

With these notations, the estimator $\hat{s}_m$ becomes

$$\hat{s}_m = \sum_{\substack{I \in \mathcal{I} \\ J \in \mathcal{J}_I}} \frac{N(I \times J)}{M(I)\mu(J)} 1_{I \times J}.$$

We define

$$\begin{aligned} \bar{s}_m &= \sum_{\substack{I \in \mathcal{I} \\ J \in \mathcal{J}_I}} \frac{\mathbb{E}[N(I \times J)]}{\mathbb{E}[M(I)]\mu(J)} 1_{I \times J}, \\ s^\star_m &= \sum_{\substack{I \in \mathcal{I} \\ J \in \mathcal{J}_I}} \frac{N(I \times J)}{\mathbb{E}[M(I)]\mu(J)} 1_{I \times J}. \end{aligned}$$

We use the triangular inequality to get

$$\mathbb{E}\left[h^2(s, \hat{s}_m)\right] \leq 2h^2(s, \bar{s}_m) + 4\mathbb{E}\left[h^2(\bar{s}_m, s^\star_m)\right] + 4\mathbb{E}\left[h^2(s^\star_m, \hat{s}_m)\right]. \tag{17}$$

It remains to control both of the three terms appearing in the right-hand side of the above inequality. The first term can be upper bounded thanks to Lemma 2 of Baraud and Birgé (2009):

$$h^2(s, \bar{s}_m) \leq 2h^2(s, V_m). \tag{18}$$

Now,

$$\begin{aligned} h^2(\bar{s}_m, s^\star_m) &= \frac{1}{2n} \sum_{\substack{I \in \mathcal{I} \\ J \in \mathcal{J}_I}} \left(\sqrt{\frac{\mathbb{E}[N(I \times J)]}{\mathbb{E}[M(I)]\mu(J)}} - \sqrt{\frac{N(I \times J)}{\mathbb{E}[M(I)]\mu(J)}} \right)^2 \mathbb{E}[M(I)]\mu(J) \\ &= \frac{1}{2n} \sum_{\substack{I \in \mathcal{I} \\ J \in \mathcal{J}_I}} \left(\sqrt{\mathbb{E}[N(I \times J)]} - \sqrt{N(I \times J)} \right)^2 \\ &\leq \frac{1}{2n} \sum_{\substack{I \in \mathcal{I} \\ J \in \mathcal{J}_I}} \frac{\left(\mathbb{E}[N(I \times J)] - N(I \times J)\right)^2}{\mathbb{E}[N(I \times J)]}. \end{aligned}$$

By taking the expectation of both sides,

$$\mathbb{E}\left[h^2(\bar{s}_m, s_m^\star)\right] \leq \frac{1}{2n} \sum_{\substack{I\in\mathcal{I}\\ J\in\mathcal{J}_I}} \frac{\mathrm{var}\left[N(I\times J)\right]}{\mathbb{E}[N(I\times J)]} \leq \frac{1}{2n} \sum_{\substack{I\in\mathcal{I}\\ J\in\mathcal{J}_I}} 1 \leq \frac{|m|}{2n}. \tag{19}$$

As to the third term,

$$\begin{aligned} h^2(\hat{s}_m, s_m^\star) &= \frac{1}{2n} \sum_{\substack{I\in\mathcal{I}\\ J\in\mathcal{J}_I}} \left(\sqrt{\frac{N(I\times J)}{M(I)\mu(J)}} - \sqrt{\frac{N(I\times J)}{\mathbb{E}[M(I)]\mu(J)}}\right)^2 \mathbb{E}[M(I)]\mu(J) \\ &= \frac{1}{2n} \sum_{\substack{I\in\mathcal{I}\\ J\in\mathcal{J}_I}} \left(\sqrt{\mathbb{E}[M(I)]} - \sqrt{M(I)}\right)^2 \frac{N(I\times J)}{M(I)} \\ &= \frac{1}{2n} \sum_{I\in\mathcal{I}} \left(\sqrt{\mathbb{E}[M(I)]} - \sqrt{M(I)}\right)^2 \sum_{J\in\mathcal{J}_I} \frac{N(I\times J)}{M(I)} \\ &\leq \frac{1}{2n} \sum_{I\in\mathcal{I}} |\mathcal{J}_I| \left(\sqrt{\mathbb{E}[M(I)]} - \sqrt{M(I)}\right)^2. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[h^2(\hat{s}_m, s_m^\star)\right] &\leq \frac{1}{2n} \sum_{I\in\mathcal{I}} |\mathcal{J}_I|\, \mathbb{E}\left[\frac{(M(I)-\mathbb{E}[M(I)])^2}{\mathbb{E}[M(I)]}\right] \\ &\leq \frac{1}{2n} \sum_{I\in\mathcal{I}} |\mathcal{J}_I| \frac{\mathrm{var}[M(I)]}{\mathbb{E}[M(I)]} \\ &\leq \frac{1}{2n} \sum_{I\in\mathcal{I}} |\mathcal{J}_I| \\ &\leq \frac{|m|}{2n}. \end{aligned} \tag{20}$$

Gathering (17), (18), (19) and (20) leads to the result. □

4.4. Proof of Proposition 4. We use in this proof the notations introduced in the proof of Proposition 3. We derive from the triangular inequality and (18),

$$h^2(s, \hat{s}_m) \leq 4h^2(s, V_m) + 4h^2(\bar{s}_m, s_m^\star) + 4h^2(s_m^\star, \hat{s}_m), \tag{21}$$

and it remains to bound the two last terms from above. Yet,

$$h^2(\bar{s}_m, s_m^\star) = \frac{1}{2n} \sum_{\substack{I\in\mathcal{I}\\ J\in\mathcal{J}_I}} \left(\sqrt{\mathbb{E}[N(I\times J)]} - \sqrt{N(I\times J)}\right)^2.$$

Theorem 8 of Baraud and Birgé (2009) (applied with $A = 1$ and $\kappa = 1$) shows that, for all $x > 0$, with probability larger than $1 - e^{-x}$,

$$\sum_{\substack{I\in\mathcal{I}\\ J\in\mathcal{J}_I}} \left(\sqrt{\mathbb{E}[N(I\times J)]} - \sqrt{N(I\times J)}\right)^2 \leq 8|m| + 202x.$$

Now,

$$\begin{aligned} h^2(\hat{s}_m, s_m^\star) &= \frac{1}{2n} \sum_{\substack{I\in\mathcal{I}\\ J\in\mathcal{J}_I}} \left(\sqrt{\frac{N(I\times J)}{M(I)\mu(J)}} - \sqrt{\frac{N(I\times J)}{\mathbb{E}[M(I)\mu(J)]}} \right)^2 \mathbb{E}[M(I)]\mu(J) \\ &= \frac{1}{2n} \sum_{\substack{I\in\mathcal{I}\\ J\in\mathcal{J}_I}} \left(\sqrt{\mathbb{E}[M(I)]} - \sqrt{M(I)} \right)^2 \frac{N(I\times J)}{M(I)} \\ &= \frac{1}{2n} \sum_{I\in\mathcal{I}} \left(\sqrt{\mathbb{E}[M(I)]} - \sqrt{M(I)} \right)^2 \sum_{J\in\mathcal{J}_I} \frac{N(I\times J)}{M(I)} \\ &\le \frac{1}{2n} \sum_{I\in\mathcal{I}} \left(\sqrt{\mathbb{E}[M(I)]} - \sqrt{M(I)} \right)^2. \end{aligned}$$

A new application of Theorem 8 of Baraud and Birgé (2009) shows that for all $x > 0$, with probability larger than $1 - e^{-x}$,

$$\sum_{I\in\mathcal{I}} \left(\sqrt{\mathbb{E}[M(I)]} - \sqrt{M(I)} \right)^2 \le 8|\mathcal{I}| + 202x.$$

We then deduce from (21) that for all $x > 0$, with probability larger than $1 - 2e^{-x}$,

$$h^2(s, \hat{s}_m) \le 4h^2(s, V_m) + 16\frac{|m| + |\mathcal{I}|}{n} + 808\frac{x}{n}$$

The result follows with $x = n\xi + \log 2$. □

4.5. Proof of Theorem 5. Let, for each model $V \in \mathbb{V}$, T_V be a subset of V satisfying the two following conditions:

- for all $g \in V$, there exists $f \in T_V$ such that $d_2(f, g) \le 1/n$
-

$$|\{f \in T_V,\, d_2(f, 0) \le 2\}| \le (4n + 1)^{\dim V}.$$

For instance, we can define T_V as a maximal $1/n$-separated subset of V in the metric space $(\mathbb{L}^2(A, \nu \otimes \mu), d_2)$ in the sense of Definition 5 of Birgé (2006). The bound on the cardinality is then given by Lemma 4 of Birgé (2006). Let

$$S_V = \left\{ f_+^2,\ f \in T_V,\, d_2(f, 0) \le 2 \right\} \cup \{0\} \quad \text{and} \quad S = \bigcup_{V\in\mathbb{V}} S_V.$$

We now define the map $\bar{\Delta}$ on S by

$$\bar{\Delta}(f) = \inf_{\substack{V\in\mathbb{V}\\ S_V \ni f}} \left\{ \Delta(V) + \log|S_V| \right\}.$$

Without loss of generality, we can assume that $S \subset \mathcal{L}(A, \mu)$ (thanks to Proposition 2). Theorem 1 shows that there exists an estimator $\hat{s}$ satisfying for all $\xi > 0$ and probability larger than $1 - e^{-n\xi}$,

$$h^2(s, \hat{s}) \le c_1 \inf_{f\in S} \left\{ h^2(s, f) + L\frac{\bar{\Delta}(f)}{n} \right\} + c_2\xi.$$

Hence,

$$\begin{aligned} h^2(s,\hat{s}) &\leq c_1 \inf_{V\in\mathbb{V}} \left\{ h^2(s,S_V) + L\frac{\Delta(V)+\log|S_V|}{n} \right\} + c_2\xi \\ &\leq c_1 \inf_{V\in\mathbb{V}} \left\{ \kappa \inf_{\substack{f\in T_V \\ d_2(f,0)\leq 2}} d_2^2(\sqrt{s},f) + L\frac{\Delta(V)+\log|S_V|}{n} \right\} + c_2\xi. \end{aligned}$$

As $d_2(\sqrt{s},0)\leq 1$ and $0\in S_V$,

$$\begin{aligned} \inf_{\substack{f\in T_V \\ d_2(f,0)\leq 2}} d_2(\sqrt{s},f) &= \inf_{f\in T_V} d_2(\sqrt{s},f) \\ &\leq \inf_{f\in V} d_2(\sqrt{s},f) + 1/n. \end{aligned}$$

Therefore,

$$\begin{aligned} h^2(s,\hat{s}) &\leq c_1 \inf_{V\in\mathbb{V}} \left\{ 2\kappa d_2^2(\sqrt{s},V) + 2\frac{\kappa}{n^2} + L\frac{\Delta(V)+\log\left(1+(4n+1)^{\dim V}\right)}{n} \right\} + c_2\xi \\ &\leq C\left(\inf_{V\in\mathbb{V}} \left\{ \kappa d_2^2(\sqrt{s},V) + \frac{\Delta(V)+(\dim V)\log n}{n} \right\} + \frac{\kappa}{n^2} + \xi \right) \end{aligned}$$

for C large enough. □

4.6. Structural assumptions. Theorem 2 and Corollary 1 of Baraud and Birgé (2014) are useful tools to deal with structural assumptions. They show how to build collections $\mathbb{V}$ of linear spaces V with good approximation properties with respect to composite functions f of the form $f = g\circ u$. Using these results is the strategy of Sart (2014) to get bounds on $\varepsilon_{\mathscr{F}}(f)$ for classes $\mathscr{F}$ corresponding to structural assumptions on s. Nevertheless, this direct application of the results of Baraud and Birgé (2014) (with $\tau = \log n/n$) leads to an unnecessary additional logarithmic term in the risk bounds. A careful look at the proof of Theorem 2 of Baraud and Birgé (2014) shows that the following result holds.

Theorem 6. *Suppose that Assumption 1 holds and that $\nu\otimes\mu(A) = 1$. Let $l\in\mathbb{N}^\star$ and $\mathcal{L}^\infty([0,1]^l)$ be the set of bounded functions on $[0,1]^l$ endowed with the supremum distance*

$$d_\infty(g_1,g_2) = \sup_{x\in[0,1]^l} |g_2(x)-g_1(x)| \quad \text{for } g_1,g_2\in\mathcal{L}^\infty([0,1]^l).$$

Let $\mathcal{U}$ be the set of functions $u=(u_1,\dots,u_l)$ going from A to $[0,1]^l$ and

$$\mathscr{F} = \left\{ g\circ u,\ g\in \bigcup_{\boldsymbol{\alpha}\in(0,1]^l} \mathcal{H}^{\boldsymbol{\alpha}}([0,1]^l),\ u\in\mathcal{U} \right\}.$$

Let $\mathbb{F}$ be an at most countable collection of finite dimensional linear subspaces F of $\mathcal{L}^\infty([0,1]^l)$ endowed with a map $\Delta_{\mathbb{F}}\geq 1$ satisfying

$$\sum_{F\in\mathbb{F}} e^{-\Delta_{\mathbb{F}}(F)} \leq 1.$$

Let, for all $j\in\{1,\dots,l\}$, $\mathbb{T}_j$ be an at most countable collection of subsets T of $\mathbb{L}^2(A,\nu\otimes\mu)$. We assume that each T is either a unit set, or a finite dimensional linear space. If T is a singleton,

we set $\dim T = 0$. *If* T *is a non trivial linear space,* $\dim T$ *stands for its usual linear dimension. We endow* $\mathbb{T}_j$ *with a non-negative map* $\Delta_{\mathbb{T}_j}$ *satisfying*

$$\sum_{T\in\mathbb{T}_j} e^{-\Delta_{\mathbb{T}_j}(T)} \leq 1.$$

Then, there exist a collection $\mathbb{V}$ *and a map* Δ *such that for all function* $f \in \mathscr{F}$ *of the form* $f = g \circ u$, *with* $g \in \mathcal{H}^{\boldsymbol{\alpha}}([0,1]^l)$ *for some* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l) \in (0,1]^l$, *and* $u = (u_1, \ldots, u_l) \in \mathcal{U}$,

$$\begin{aligned} C\varepsilon_{\mathscr{F}}(f) \quad \leq \quad & \sum_{j=1}^{l} \inf_{T\in\mathbb{T}_j} \left\{ l|g_j|^2_{\alpha_j} d_2^{2\alpha_j}(u_j, T) + \frac{\Delta_{\mathbb{T}_j}(T) + (\dim T)\mathcal{L}_{j,T}}{n} \right\} \\ & + \inf_{F\in\mathbb{F}} \left\{ d^2_\infty(g, F) + \frac{\Delta_{\mathbb{F}}(F) + (\dim F)\log n}{n} \right\}. \end{aligned}$$

In the above inequality, C *is a positive universal constant,* g_j, $|g_j|_{\alpha_j}$ *are defined as explained in Section 2.5.1, and* $\mathcal{L}_{j,T}$ *is defined when* $\dim T > 0$ *by*

$$\begin{aligned} \mathcal{L}_{j,T} \quad &= \quad \left[\alpha_j^{-1} \log\left(nl|g_j|^2_{\alpha_j}/\dim T\right)\right] \vee 1 \\ &\leq \quad C' \left[\log n \vee \log\left(|g_j|^2_{\alpha_j}/\dim T\right) \vee 1\right] \end{aligned}$$

for C' *depending only on* l *and* α_j. *When* $\dim T = 0$, $\mathcal{L}_{j,T} = 1$.

The proof of (9) is almost the same as the one of Corollary 4 of Sart (2014). The only difference is that we apply the above theorem in place of Theorem 2 of Baraud and Birgé (2014) with $\tau = \log n/n$.

We now turn to the proof of (10). Note that a function f of the form $f(x,y) = g(<\theta, x>, y)$ can be rewritten as $f(x,y) = g(u_1(x,y), u_2(x,y), \ldots, u_{1+d_2}(x,y))$ where $u_1(x,y) =<\theta, x>$ and $u_j(x,y) = y$ for $j \in \{2, \ldots, 1+d_2\}$. There exists a pair $(\mathbb{F}, \Delta_{\mathbb{F}})$ such that for all $\boldsymbol{\alpha} \in (0,+\infty)^{1+d_2}$, $g \in \mathcal{H}^{\boldsymbol{\alpha}}([0,1]^{1+d_2})$,

$$\inf_{F\in\mathbb{F}} \left\{ d^2_\infty(g, F) + \frac{\Delta_{\mathbb{F}}(F) + (\dim F)\log n}{n} \right\} \leq C_1 \left[|g|^{\frac{2(1+d_2)}{1+d_2+2\bar{\boldsymbol{\alpha}}}}_{\boldsymbol{\alpha},\infty} \left(\frac{\log n}{n}\right)^{\frac{2\bar{\boldsymbol{\alpha}}}{2\bar{\boldsymbol{\alpha}}+1+d_2}} + \frac{\log n}{n} \right]$$

for a constant C_1 depending only on d_2, $\boldsymbol{\alpha}$ (see, for instance, Baraud and Birgé (2014)). Let for $\theta \in \mathbb{R}^{d_1}$, u_θ be the function defined by $u_\theta(x,y) =< \theta, x >$ and T_1 be the linear space defined by $T_1 = \{u_\theta, \theta \in \mathbb{R}^{d_1}\}$. We use the above theorem with $l = 1 + d_2$, $\mathbb{T}_1 = \{T\}$, $\Delta_{\mathbb{T}_1}(T) = 1$, $\mathbb{T}_j = \{\{u_j\}\}$, $\Delta_{\mathbb{T}_j}(\{u_j\}) = 0$ for $j \in \{2, \ldots, 1+d_2\}$ to derive that for all $\boldsymbol{\alpha} \in (0,+\infty)^{1+d_2}$, $g \in \mathcal{H}^{\boldsymbol{\alpha}}([0,1]^{1+d_2})$, $\theta \in \mathcal{B}_1(0,1)$, and all function $f \in \mathscr{F}$ of the form $f(x,y) = g(<\theta, x>, y)$,

$$\varepsilon_{\mathscr{F}}(f) \leq C_1 \left[|g|^{\frac{2(1+d_2)}{1+d_2+2\bar{\boldsymbol{\alpha}}}}_{\boldsymbol{\alpha}} \left(\frac{\log n}{n}\right)^{\frac{2\bar{\boldsymbol{\alpha}}}{2\bar{\boldsymbol{\alpha}}+1+d_2}} + \frac{\log n}{n} \right] + C_2 d_1 \frac{\log n \vee \log\left(|g|^2_{\alpha_1\wedge 1}/d_1\right)}{n}$$

where C_1 depends only on d_2, $\boldsymbol{\alpha}$ and C_2 depends only on $\alpha_1 \wedge 1$, d_2. □

References

Akakpo, N. (2012). Adaptation to anisotropy and inhomogeneity via dyadic piecewise polynomial selection. *Mathematical Methods of Statistics*, 21:1–28.

Akakpo, N. and Lacour, C. (2011). Inhomogeneous and anisotropic conditional density estimation from dependent data. *Electronic Journal of Statistics*, 5:1618–1653.

Baraud, Y. (2011). Estimator selection with respect to Hellinger-type risks. *Probability Theory and Related Fields*, 151(1-2):353–401.

Baraud, Y. (2013). Estimation of the density of a determinantal process. *Confluentes Mathematici*, 5(1):3–21.

Baraud, Y. and Birgé, L. (2009). Estimating the intensity of a random measure by histogram type estimators. *Probability Theory and Related Fields*, 143:239–284.

Baraud, Y., Birgé, L., and Sart, M. (2016). A new method for estimation and model selection: ρ-estimation. *Inventiones mathematicae*, pages 1–93.

Baraud, Y. and Birgé, L. (2014). Estimating composite functions by model selection. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(1):285–314.

Bertin, K., Lacour, C., and Rivoirard, V. (2013). Adaptive estimation of conditional density function. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*. To appear.

Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Annales de l'Institut Henri Poincaré. Probabilités et Statistique*, 42(3):273–325.

Birgé, L. (2007). Model selection for Poisson processes. In *Asymptotics: particles, processes and inverse problems*, volume 55 of *IMS Lecture Notes Monogr. Ser.*, pages 32–64. Inst. Math. Statist., Beachwood, OH.

Birgé, L. (2012). Robust tests for model selection. In *From Probability to Statistics and Back: High-Dimensional Models and Processes. A Festschrift in Honor of Jon Wellner*, volume 9, pages 47–64. IMS Collections.

Bott, A.-K. and Kohler, M. (2015). Adaptive estimation of a conditional density. *International Statistical Review*.

Brunel, E., Comte, F., and Lacour, C. (2007). Adaptive estimation of the conditional density in the presence of censoring. *Sankhyā: The Indian Journal of Statistics*, pages 734–763.

Chagny, G. (2013). Warped bases for conditional density estimation. *Mathematical Methods of Statistics*, 22(4):253–282.

Cohen, S. and Le Pennec, E. (2011). Conditional Density Estimation by Penalized Likelihood Model Selection and Applications. *ArXiv e-prints*.

De Gooijer, J. G. and Zerom, D. (2003). On conditional density estimation. *Statistica Neerlandica*, 57(2):159–176.

Devroye, L. and Lugosi, G. (1996). A universally acceptable smoothing factor for kernel density estimates. *The Annals of Statistics*, pages 2499–2512.

Efromovich, S. (2007). Conditional density estimation in a regression setting. *The Annals of Statistics*, 35(6):2504–2535.

Fan, J., Peng, L., Yao, Q., and Zhang, W. (2009). Approximating conditional density functions using dimension reduction. *Acta Math. Appl. Sin. Engl. Ser.*, 25(3):445–456.

Fan, J., Yao, Q., and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206.

Fan, J. and Yim, T. (2004). A data-driven method for estimating conditional densities. *Biometrika*, 91:819–834.

Faugeras, O. P. (2009). A quantile-copula approach to conditional density estimation. *Journal of Multivariate Analysis*, 100(9):2083–2099.

Györfi, L. and Kohler, M. (2007). Nonparametric estimation of conditional distributions. *IEEE Transactions on Information Theory*, 53(5):1872.

Hall, P. and Yao, Q. (2005). Approximating conditional distribution functions using dimension reduction. *The Annals of Statistics*, 33(3):1404–1421.

Hyndman, R. J., Bashtannyk, D. M., and Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5(4):315–336.

Hyndman, R. J. and Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *Journal of nonparametric statistics*, 14(3):259–278.

Massart, P. (2003). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer Berlin/Heidelberg. École d'été de Probabilités de Saint-Flour.

Rosenblatt, M. (1969). Conditional probability density and regression estimators. In *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968)*, pages 25–31. Academic Press, New York.

Sart, M. (2014). Estimation of the transition density of a Markov chain. *Annales de l'Institut Henri Poincaré. Probabilités et Statistique*, 50(3):1028–1068.

Sart, M. (2015). Model selection for poisson processes with covariates. *ESAIM: Probability and Statistics*, 19:204–235.

Univ Lyon, UJM-Saint-Etienne, CNRS, Institut Camille Jordan UMR 5208, F-42023, SAINT-ETIENNE, France

E-mail address: mathieu.sart@univ-st-etienne.fr